Informatics in Medicine Unlocked 16 (2019) 100194

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

An automated workflow for MALDI-ToF mass spectra pattern identification on large data sets: An application to detect aneuploidies from pregnancy urine

Ricardo J. Pais[a], R. Zmuidinaite[a], S.A. Butler[a], R.K. Iles[a,b,*]

[a] MAP Sciences Ltd, Priory Business Park, Bedford I-lab Stannard Way, Bedford, MK44 3RZ, UK
[b] Dean's Office, College of Healthcare Sciences, Abu Dhabi University, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Keywords:
MALDI-ToF
Pattern recognition
Quality control
Comparative intensity data
Automated processing

ABSTRACT

Urine from first trimester pregnancies has been found to be rich in information related to aneuploidies and other clinical conditions. Mass spectral analysis derived from matrix assisted laser desorption ionization (MALDI) time of flight (ToF) data has been proven to be a cost effective method for clinical diagnostics. However, urine mass spectra are complex and require data modelling frameworks. Therefore, computational approaches that systematically analyse big data generated from MALDI-ToF mass spectra are essential. To address this issue, we developed an automated workflow that successfully processed large data sets from MALDI-ToF which is 100-fold faster than using a common software tool. Our method performs accurate data quality control decisions, and generates a comparative analysis to extract peak intensity patterns from a data set. We successfully applied our framework to the identification of peak intensity patterns for Trisomy 21 and Trisomy 18 gestations on data sets from maternal pregnancy urines obtained in the UK and China. The results from our automated comparative analysis have shown characteristic patterns associated with aneuploidies in the first trimester pregnancy. Moreover, we have shown that the intensity patterns depended on the population origin, gestational age, and MALDI-ToF instrument.

1. Introduction

Matrix assisted laser desorption ionization (MALDI) time of flight (ToF) mass spectrometry (MS) is considered a direct, rapid and affordable analytical method in clinical microbiology [1–3]. It is a technique which involves measurements of mass-to-charge ratios (m/z) of the ions formed when a sample is ionized. This technology has been further demonstrated to provide a substantial reduction in the costs per test associated with diagnosis, whilst still keeping high diagnostic effectiveness [4,5]. The application of MALDI-ToF in urine is a non-invasive approach particularly suitable for the development of mass market diagnostic or screening tests in global populations, with particular relevance to obstetrics and reproductive medicine [6,7].

Urine is a complex mixture of proteins, peptides, sugar-based moieties, lipids and metabolic by-products [8–10]. Hence, urine MS generated by MALDI-ToF are too complex to be tackled with conventional statistical methods and biomarker identification [10,11]. Moreover, MS of urines generated by MALDI-ToF often produce raw data spectra with complex baselines that require several data processing steps, using more sophisticated computational methods, to extract meaningful information [11]. Several processing tools are already available in free software such as mMass, where the user can visualize, crop, zoom, smooth, perform baseline correction and export the data of a single spectrum in several available formats [12,13]. However, this consumes time, resources and is prone to human bias; in particular when dealing with big data coming from different labs. Therefore, automated frameworks are required to deal with such large data sets and the complex nature of biological samples.

Recently, the application of MALDI-ToF to clinical diagnostics in urines has started to show promising results [6,14–18]. We have previously examined spectral patterns from 2000 to 14000 m/z, which incorporated the major urinary metabolite of human chorionic gonadotropin (hCG) – beta core fragment hCG (hCGβcf), and this study suggested the existence of characteristic intensity patterns associated with aneuploidies in pregnancy [18]. Moreover, we further identified gestational age differences across spectra from urine samples during first trimester pregnancy together with other confounding variables [19]. These results were achieved through several months of

* Corresponding author. MAP Sciences Ltd, Priory Business Park, Bedford I-lab Stannard Way, Bedford, MK44 3RZ, UK.
E-mail address: Ray.iles@MAPSciences.com (R.K. Iles).

https://doi.org/10.1016/j.imu.2019.100194
Received 1 May 2019; Received in revised form 22 May 2019; Accepted 25 May 2019
Available online 29 May 2019

processing, analysing and manually curating of data.

For the development of mass market screening tests, with acceptable diagnostic power, larger data sets need to be analysed (> 10000 samples). To deal with such demand, we have developed an automated computational workflow that processes the raw data from MALDI-ToF, performs quality control decisions, and systematically performs a comparative analysis to identify peak intensity patterns on a given data subset. We have specifically designed this automated workflow to deal with MS data from pregnancy urine, and tested its performance against large data sets (100–10000 data files). Here, we show the application of our automated workflow and computational methods for the identification of MS patterns associated to Trisomy 21 (T21), Trisomy 18 (T18) and preeclampsia in the first trimester of pregnancy. Furthermore, we have also analysed ethnic origin of the cohort, gestational age and any variability from the use of different MALDI instruments as possible confounding variables. Our approach is not based on the identification of biomarkers, but instead on spectral profiling and pattern recognition to predict pregnancy outcomes.

2. Methods

2.1. Data used for this study

For this study, we have used raw mass spectral data produced from cohorts of pregnancy urines from UK and Chinese populations. Mass spectral data was produced with two different MALDI instruments: a Shimadzu Axima model CFR + (Axima) and Skyray (China) MALDI-ToFmodel MicroTyper MS (Skyray). Samples data from UK populations was collected at the Harris Birthright Centre, The Fetal Medicine Research Institute, King's College Hospital and analysed in MAP Sciences' laboratory in Bedford UK. Chinese population pregnancy urine samples were collected at the Nanjing Maternity and Child Health Care Hospital, in Nanjing China and mass spectral data was generated by KeyGEN BioTECH Co Ltd in Nanjing China.

Raw data files were converted from mzml into comma-separated text files containing *m/z* values and signal intensity using the MALDIquant software package [20]. We constructed files as comma-separated values which cross referenced to clinical information provided about our cohorts e.g. sample file names, IDs, clinical outcomes and gestational age (metadata files). With these metadata files, we defined data sets organized by population origin (China or UK) and MALDI instrument used (Axima or Skyray), (Table 1). For the pattern identification analysis, we only considered spectral data that passed quality control and which had known clinical outcomes & gestational age in their associated metadata (Table 1). For tested groups, we gathered in the above-mentioned data sets, all data from pregnancies with trisomy 21 or 18. In addition, we also added data from pregnancy with preeclampsia without any other associated clinical observation. As control groups, we selected data from singleton pregnancy, with no medication usage by the mother, and without any known diseases associated to either baby or mother. For the control group, the representative 200 samples for each gestational age group (weeks 11, 12, 13 and 14) was defined to allow faster analysis and minimize bias.

2.2. Ethics approval

This study was approved by King's College Hospital Ethics Committee (02-03-033). And by the internal Ethics committee of MAP Sciences, where all samples were received anonymised for analysis. Written informed consent was obtained from all women agreeing to participate in the study and held at King's College Hospital.

2.3. Pre-processing and quality control

All mass spectra raw data files were systematically pre-processed and spectral quality evaluated. For this purpose, we have developed a computational framework which systematically picks up each file in a folder and: 1) Performs a baseline correction on the spectral data; 2) Normalises the intensity of each *m/z* value with the total intensity of the spectra; 3) Estimates the local noise (LN) within a 5 point interval on *m/z* axis based on the method proposed in Ref. [21]; 4) Identifies putative, well resolved hCGβcf peaks as a quality control measure in the context of pregnancy [15]; 5) Evaluates the quality of the spectra based on the signal to local noise ratio (SLNR). In this workflow, the spectral baseline was automatically estimated by fitting a 5-degree polynomial curve, subtracted from the raw data, and removing possible negative values. The local noise of each raw data sample was estimated using the maximum absolute difference in the amplitude of the all possible slopes with a sliding window of 5 points intervals [21]. The key pregnancy protein hCGβcf was identified by iteratively searching within 9600–10000 m/z range for a well resolved peak at 9750 ± 25 m/z [15]. Here, we have used a generic peak picking algorithm in combination with data smoothing using Savitzky-Golay filter [22,23]. Samples were only accepted if the SLNR for at least one peak within 9600–10000 m/z range was higher than 5-fold. Pre-processed data files were automatically exported to a folder and associated to meta data information. All samples were also manually checked for erroneous acceptance or rejection by running the script with an optional plot of raw data spectra, fitted baseline, pre-processed spectra and peaks detected. For this purpose, we randomly selected a set of data to perform the manually pre-processing option implemented in the script. All data was also manually checked for quality by plotting each data file in mMass software and performing a visual inspection of noise and presence of hCGβcf peak.

2.4. Spectral features extraction

Spectral features composed by peaks positions, normalized peaks intensity, and correspondent error due to noise were automatically extracted from pre-processed data samples. On each pre-processed MS data, we iteratively searched for well resolved peaks in the range of 2000–13000 m/z using a modified peak finding algorithm in combination with data smoothing using Savitzky-Golay filter [22,23]. The reason for this choice was to remove enough noise to allow good peak detection, keeping the asymmetry properties of peak signals in urine MS and not introduce artefacts by forcing a gaussian distribution on peaks which are not symmetric. The parameters of Savitzky -Golay filter (window size of 10 m/z and 5 cycles) were previously optimized for the urine MS data from MALDI to remove noise and provide a reasonable estimate of the position of the peaks in raw data using the python peakutils function. The values of intensities for all peaks were subsequently computed by getting the correspondent values in the raw data to avoid deviations.

Only the peaks with an identified signal to noise ratio greater than 5-fold were selected. The error associated to each peak was estimated based on the average LN around the peak region. Finally, a file containing all spectral features was generated. In this method, we have also included a spectral alignment that makes all peak positions shift according to a reference peak to minimize *m/z* errors in peak positions as an optional feature. Briefly, when a reference peak is known, the computational method searches within a 50 m/z window and computes the difference between the reference peak mass and the actual mass of an observed peak found in that window. Peak position values are then corrected throughout the spectra by applying the computed difference and shifting peak positions correspondingly.

2.5. Peak enrichment and intensity differences analysis

Peak enrichment and the differences in intensity between a tested data set (clinical outcome) and a control group (control data set) were computed using an automated computational methodology. As inputs, the method takes the file containing peak positions, peak intensities and

Table 1
Data sets used in the study and characteristics. Outcome abbreviations: T21 is trisomy 21; T18 is trisomy 18; normal is non-aneuploid and without any other known clinical diagnosis; other is a sample with diagnosis such as anaemia, hyperlipidaemia, diabetes; high risk is a non-aneuploid associated with a high risk for aneuploidies without any other known clinical diagnosis.

Data sets	MALDI Machine	Origin (population)	Gestational age (weeks)	Number of samples and outcomes	Objective
Axima	Axima	China	All	6 T21 2 T18 64 Preeclampsia 2020 Normal 980 Other	QC validation
Skyray	Skyray	China	All	6 T21 2 T18 64 Preeclampsia 2100 Normal 980 Other	QC validation
ChAx	Axima	China	11–14th	3 T21 2 T18 28 Preeclampsia 789 Normal	Pattern identification
13W ChAx	Axima	China	13th	1 T21 1 T18 5 Preeclampsia 200 Normal	Pattern identification
ChSky	Skyray	China	11–14th	2 T21 2 T18 28 Preeclampsia 743 Normal	Pattern identification
ChSky	Skyray	China	13th	1 T21 1 T18 7 Preeclampsia 200 Normal	Pattern identification
UK	Axima	United Kingdom	11–14th	20 T21 3 T18 74 High risk	Pattern identification
13W UK	Axima	United Kingdom	13th	12 T21 1 T18 47 High risk	Pattern identification

the estimated relative errors for all samples, ID numbers, outcomes and gestational ages in a list. The data is then separated into two groups (tested and control) according to clinical outcomes and then further split into classes with different weeks of gestational age. For each group and classes, peak enrichments were computed between 2000 and 13000 m/z with constant 50 m/z interval. Peak enrichments (E_K) were iteratively computed according to equation (1) where: P_i and P_j are sets of values of peaks positions for a tested sample i and a control sample j, respectively; n and m are the total numbers of samples of tested and control outcomes, respectively; and K is a set of continuous values that define an *m/z* region for which enrichment is calculated. Peak enrichments were computed only considering the presence or absence of peaks in a given K region, regardless of multiple peaks in K region. Thus, we have introduced an if conditional statement in equation (1) to mathematically represent how peak enrichments were computed.

$$E_K = \frac{\sum_i^n (1\ if\ \ P_i \cap K \neq \varnothing)/n}{\sum_j^m (1\ if\ \ P_j \cap K \neq \varnothing)/m \ + \ \sum_i^n (1\ if\ \ P_i \cap K \neq \varnothing)/n} \quad (1)$$

Simultaneously, the medians of the values of intensity for each region (K) were computed for each group (tested and control). Errors associated with peak intensity were used to generate variability assuming intrinsic gaussian noise. This rendered a much better accuracy around medians for a small number of samples. Next, the differences between medians of tested and control groups were calculated and expressed in terms of their logarithm gains of intensity (LGM). All E_K and LGMs were saved in files for subsequent visualization and analysis.

2.6. Pattern identification

Peak enrichment and the intensity differences analysis were combined to identify patterns of tested data set (clinical outcome) within the control group (control data set) using a computational method design in our lab. In this analysis, the peak enrichments, medians, 10th and 90th quantiles were calculated for each group and gestational age classes. With this data, regions were selected according to a threshold of enrichment differences between tested and control groups, or with a logarithm gain equivalent of 2-fold differences in medians as an alternative. To minimize overlap between groups, we also implemented a filter for only selecting regions which have less than 10% of data shared. To classify the quality of the patterns, we implemented a tracking method for the combinatorial effects of regions in terms of the capacity to describe the two data sets (percentage of group coverage). We further programmed our computational approach to iteratively perform the analysis with a 50 m/z sliding window, whereas the selected regions and results are printed to a data file.

2.7. Computations and code

The scripts that compose the computational workflow described in this work were written in Python version 2.7. All scripts were developed in MAP Sciences bioinformatic's laboratory under Anaconda Distribution version 5.3. All scripts produced are under copyright and commercial intellectual property protection. All computations were performed under Microsoft windows 10 Pro 64-bit operating system, using a Dell machine with an Intel(R) core(TM) i7-8550U CPU @ 1.80 GHz - 2.00 GHz and 16 GB RAM. R code was produced and run under RStudio® statistical environment version 1.1 for the conversion of raw data, generation of heatmaps, cluster analysis and bar plots.

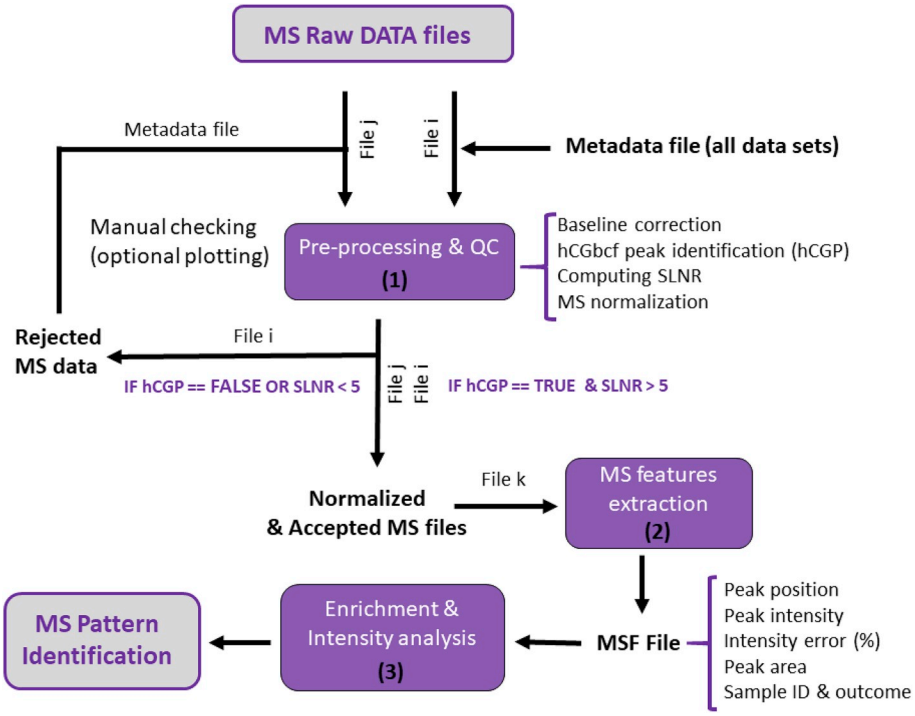

Fig. 1. Pipeline for pattern identification of large data sets of mass spectra from MALDI ToF. Data pre-processing **(1)** starts from a folder containing all stored raw data files (MS raw data files). Next, each file i is iteratively picked up and undergoes pre-processing, where an unsupervised quality control decision is made. Accepted and rejected files are stored in different folders. Quality control decision is based on the detection of hCGbeta core fragment peak (hCGP) greater than 5-fold the signal to local noise ratio (SLNR). Rejected data are re-run with the optional manual checking method, where the user checks and controls if the data was correctly rejected. Each file k in accepted folder is picked up and further processed for spectral features extraction **(2)**. The method returns one file (MSF file) containing all features extracted for all data processed. Finally, the MSF file is pickup and an enrichment analysis combined with an analysis of intensity differences on a tested and a control group contained in MSF file generates a pattern for the tested group **(3)**.

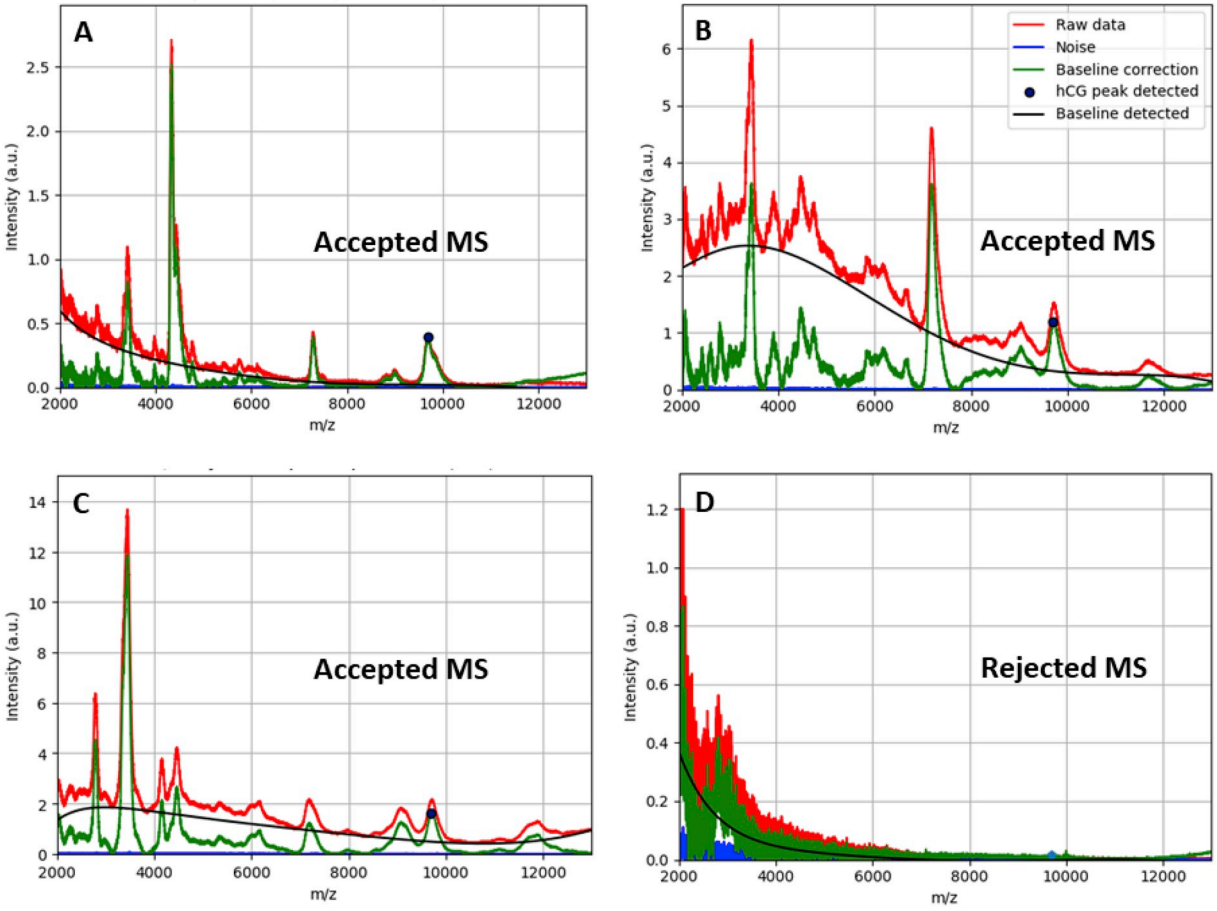

Fig. 2. Examples of automatically accepted and rejected MS from pre-processed MS Raw data of urine samples. The Illustrative examples in panel A to C were selected from a set of pre-processed raw data files to show different baseline adjustments. Example in panel D was selected to illustrate a bad sample with no hCG peak detected and high degree of noise. Plots were generated with automated pre-processing method under the manual checking optional feature.

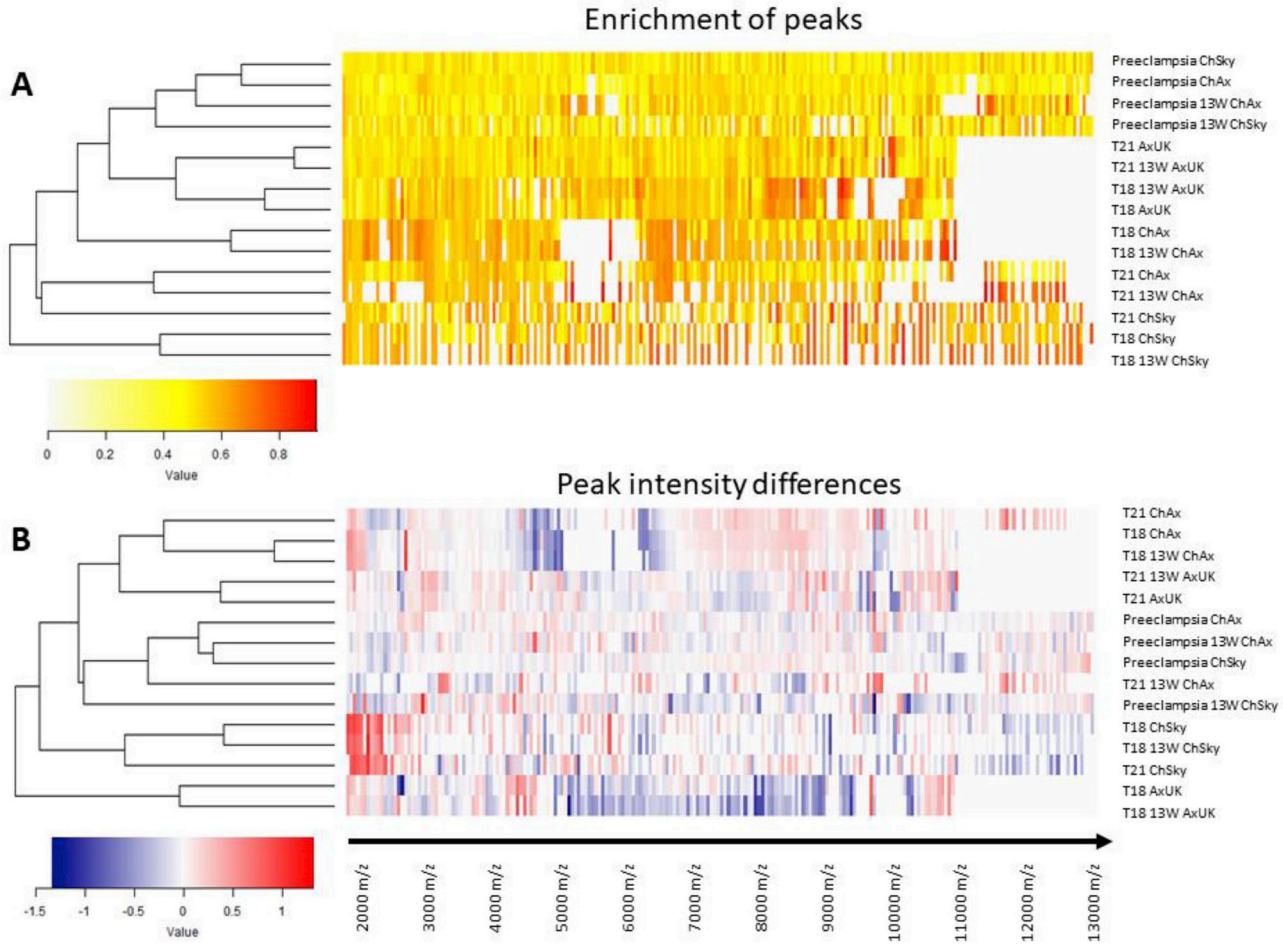

Fig. 3. Peak enrichment and intensity differences of urine mass spectra data of pregnant with aneuploidies and preeclampsia. Each clinical outcome was analysed using the automated method using the data sets in Table 1. Subsets with only the 13th week of gestation are indicated by 13W. Data sets from different MALDI machines are indicated by Sky (Skyray) and Ax (Axima). Data from China is indicated by Ch and from United Kingdom (UK). Heatmaps of peak enrichment (A) and Intensity differences (B) were analysed within 50 m/z intervals.

3. Results

3.1. Automated workflow for MS data analysis

To identify patterns in big data sets of mass spectra of urines samples form MALDI-ToF, we developed a computational workflow that automatically carries out the analysis in 3 main steps (Fig. 1). With this workflow, we first carryout a data pre-processing step, dealing with multiple raw data files one by one and also performing a quality control decision. We tested this step for computation performance and stability with small to large data sets up to 10000 data files (up to 46 GB). This resulted in a constant pre-processing rate of 1 file per second without crashing. This is in excess of 100-fold faster than the manual spectral pre-processing and file exporting using mMass. We have also included in the quality control a checking procedure (optional feature), where we re-evaluate rejected data by re-running and visually monitoring the quality of spectra and choices to minimize the loss of good data. Secondly, we run an automated extraction of the spectral features of each file from all pre-processed, accepted data. This produces a single file containing these features (MSF file). In this workflow, we included peak positions, peak intensities, estimated errors and areas between valleys as spectral features. The performance of this step was observed to be relatively fast with a constant processing rate of 41 files (MS data) per minute. On the resulting MSF file, we apply a third step, where two groups of data are systematically compared to find a pattern of peak positions and/or intensity differences for several mass ranges. This last step was considerably faster when compared to the previous steps, taking around a total of 90 s for a comparison of 3000 files processed.

3.2. Automated quality control of urine MS

The applicability of the pre-processing and quality control method on mass spectral data of pregnancy urine samples was tested with a total 6080 spectra files previously evaluated manually for its quality (see methods). As a preliminary test, we first run the workflow with the optional manual checking feature on 100 randomly selected files to visually check the quality of the baseline correction and the accuracy of detecting a well resolved hCGβcf peak. Our computational method was able to identify the hCGβcf peak and provided a good baseline correction in a complex mass spectrum such as urines (see examples in Fig. 2, panels A to C). The method was also able to correctly reject samples, where an hCGβcf peak was not visually observed and/or where the spectrum was too noisy (see example in Fig. 2, panel D). Next, we further applied the automated method for pre-processing and quality control decision with larger data sets of mass spectra of pregnancy urine samples generated from different MALDI machines (Axima and Skyray). A total of 3000 files with data from Axima and 3080 files from Skyray were pre-processed and their quality was evaluated using the unsupervised option (fully automated). According to this method, 2882 spectra from Axima and 2721 spectra from Skyray were determined to have acceptable quality and the remaining data was rejected. In addition, the quality of rejected spectra was re-evaluated

using supervised option (see Fig. 1 and methods for details) and found to have been rejected justifiably where no erroneous baseline fitting or detection of hCGβcf peak was seen. To further test the accuracy of the method, we compared the supervised quality control decisions (accept or reject) for all data with the ones previously made manually using mMass [12]. Except for two cases, the automated and unsupervised method rendered quality control decisions in agreement with the ones generated manually using mMass. On the other hand, our method accepted two spectra that were manually rejected. This was due to the presence of small hCGβcf peak, close to the defined signal to noise ratio threshold which was difficult to observe visually.

3.3. Analysis of aneuploidies from urine MS data

Here, we show the application of the developed computational workflow to urine MS pattern identification of aneuploidies in pregnancies. For this purpose, we applied our method to several data sets constructed with data from UK and China generated using two MALDI-ToF instruments from different manufacturers (Table 1). The results from the computed peak enrichment and intensity differences using the data sets are shown in Fig. 3. These results show patterns in data of T21 and T18 aneuploidies during first trimester pregnancy for different populations and MALDI-ToF instruments. In addition, we also apply our computational method in finding MS patterns in data from pregnancies with a later preeclampsia diagnosis.

Interestingly, T21 and T18 aneuploidies peak enrichment and intensity differences from UK and China were predominately clustered in the same groups, separated from the data of non-aneuploidies pregnancies with preeclampsia. This indicates that MS patterns of aneuploidies share common characteristics, which are in general distinguishable from non-aneuploidies. However, the clustering also shows that peak enrichment and intensity patterns are affected by population differences and gestational age. Nevertheless, clustering of peak enrichment and intensity differences also indicates T21 and T18 can be distinguished. Moreover, the results in Fig. 3 also pinpoints that different machines (Skyray and Axima) generate distinct patterns of peak enrichment and intensity on the same data. This indicates that MS patterns are exclusively machine specific, where the data from Skyray machine apparently results in patterns with better resolution.

Using our computational method, we identified several characteristic regions for T21 and T18 (up to 42), which combined were able to characterize 100% of the data set. For preeclampsia, we identified several regions but they could only characterize up to 71% of the data set (32 regions). In general, we found that these patterns were mainly due to intensity differences rather than peak enrichment in a particular region. The analysis using only data from 13th week of gestation rendered more identified regions in comparison with data from all weeks, indicating that gestational age is a critical variable resulting in specific marker change within pattern recognition. As illustrative examples, here we present the patterns obtained using Axima MALDI-ToF machine for UK and China in the 13th week of gestation (Fig. 4). These patterns show that T21 and T18 pregnancy urine MS have distinct characteristic regions from 2000 m/z to 13000 m/z with little overlap, showing different intensity patterns, probably reflecting the different phenotypic effects of the respective Karyotype abnormalities. Furthermore, these results also indicate that MS patterns of aneuploidies in the Chinese population are different from the ones obtained from UK, particularly in the case of T21. However, the regions between 9200 and 9400 m/z and 10300–10350 m/z of T21 MS patterns show similar intensity changes between UK and China, indicating a high degree of conservation between populations. As would be expected for a common biological condition. However, the spectral regions of differences may not be racial; as in China the study population was, not specific but, a general maternity cohort and in the UK the examined population was from pregnancies already identified as screen positive for “high risk pregnancy” by existing methodologies. For T18, the characteristic patterns were also obtained with similarity in terms of intensity changes for the regions 2000–2100 m/z and 6300–6350 m/z and again regions of differences which may reflect characteristics of “at risk’ populations versus the general pregnancy population screen.

4. Discussion

Mass spectrometry-based proteomics is a technology of growing importance in a biological and clinical research aimed at differentiating diseased and healthy states. Compared to other mass spectrometry methods, Matrix Assisted Laser Desorption Ionization Time of Flight Mass Spectrometer (MALDI-ToF MS) [24] is a very sensitive technique, where urine sample with relatively dilute protein concentration can be applied neat for direct analysis [25]. A number of authors have already demonstrated a successful utilisation of urine sample MALDI data for clinical applications such as bladder cancer diagnosis [26], differentiation between benign and malignant prostate cancer [27] and early detection of diabetic nephropathy [28]. However, these studies were limited to reduced number of samples analysed compared with the full potential of high throughput data generation of MALDI technique.

Here, we have successfully applied an automated workflow to recognize characteristic peak intensity patterns of aneuploidies from MS of urine of first trimester pregnancy. Similar MS patterns were already observed for T21 and proposed by Iles et al., which reinforces the confidence around the outputs from our automated approach [15,18]. The presented workflow would save time, human resources and facilitates in the analysis of big data sets with almost no limit of data file size. Importantly, the automated processing and analysis developed brings an advantageous breakthrough for the usage of MALDI-ToF technology as a high-throughput methodology in clinical laboratory medicine, in particular for the problem of urine analysis [10]. In addition, this method provides an unbiased quality control decision, making it suitable to be applied in routine process laboratories. This was possible due to a successful implementation of the identification of the characteristic signature of hCGβcf, the principle metabolite that represents 80–90% of the hCG material found in urine [15,29,30]. Moreover, our approach has made possible the analysis of data frames with large dimensions (over 1000), which would be computationally demanding when simply importing data into statistical software packages (i.e. R statistical package) and conducting matrix-based analysis.

Using this approach, we further extended the data analysis and identified other distinct urine MS patterns for T18 and preeclampsia in pregnant women. This opens the possibility for the development of MS pattern-based machine learning algorithms for all major fetal aneuploidy's and other clinical outcomes in prenatal screening. However, the detected MS patterns from UK and China showed population differences. This suggests that future algorithms should be designed specifically for particular populations; be it general pregnancies, high risk, or racial groupings to achieve better performances. Finally, our data analysis also pinpointed that MS patterns of urine are depended on gestational age and the MALDI-ToF instrument used to generate the data. This further suggests that the design of future detection algorithms should take into account MALDI-ToF equipment and the gestational age as confounding variables. Combined with direct and straightforward sample preparation, the workflow could be an attractive and practical method for rapid clinical applications.

In principle, the automated workflow developed in this work can be applied to all types of mass spectrometry techniques including the coupling with liquid and gas chromatography. However, our approach is particularly useful for the analysis of high throughput mass spectral data techniques such as MALDI-ToF, which is evolving towards ultra-high throughput screening [31,32]. Modern MALDI-ToF equipment can generate spectra for analysis in minutes or even seconds depending on the number of shots and laser speed [32]. In this work, the automated workflow was demonstrated to be capable of analysing data from MALDI faster than it is generated. Thus, our workflow enables to reach

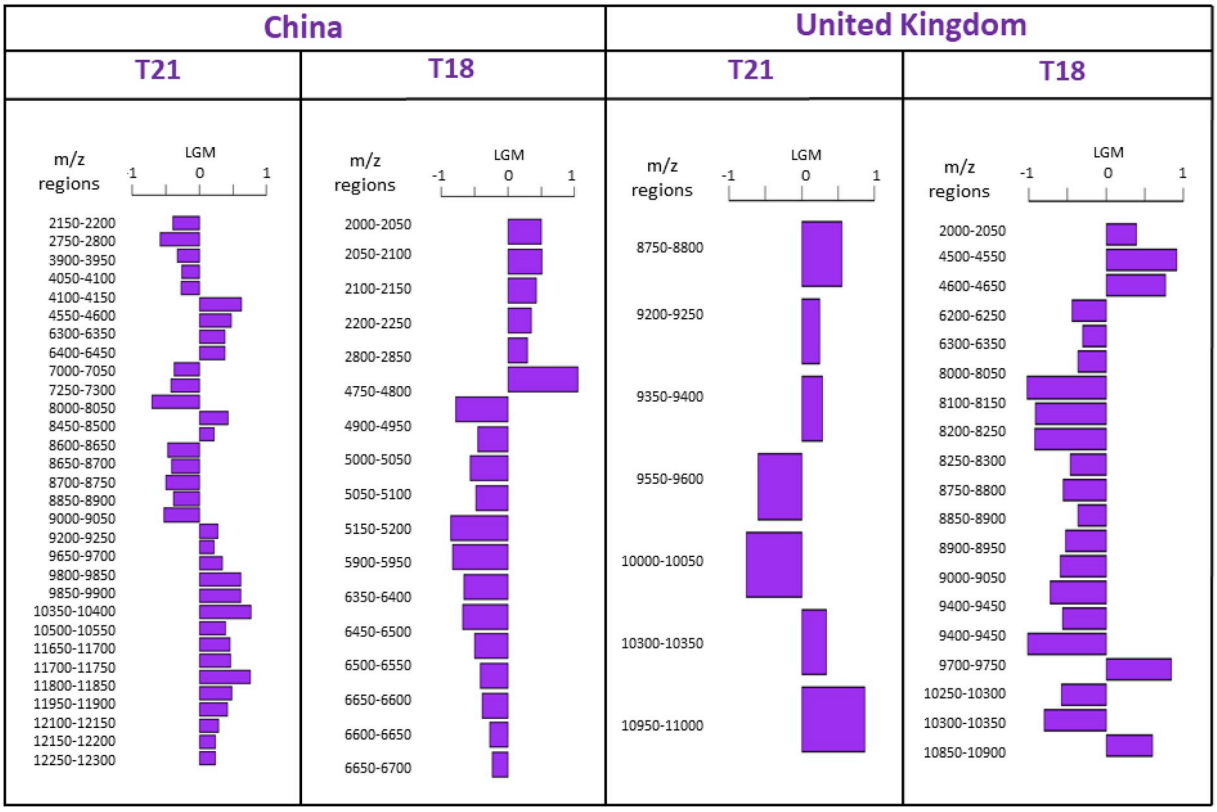

Fig. 4. MS intensity patterns of urines from pregnant with aneuploidies (T18 and T21). The patterns were obtained using the automated workflow presented in this work using the data of the 13th week of gestation from China and United Kingdom, and generated from Axima MALDI-ToF machine. Each panel contains the list of characteristic *m/z* regions and respective intensity differences in comparison with non-aneuploid pregnancy. The intensity differences are expressed in terms of the logarithm differences between medians of tested and control data sets (LGM).

the full potential of MALDI technology.

Some studies have developed other automated computational workflows to explore high throughput mass spectral data [32–35]. However, these approaches focus on biomarker discovery and quantification, rather then pattern recognition. Therefore, here we propose an alternative and innovative approach to explore mass spectral data. Although, our approach was showed to be successful for mass spectra of urines, it has also potential to be applied to other sample types such as blood, serum saliva and cell culture medium. Currently, we are adapting the workflow on embryo culture media from IVF to identify patterns that correlate with embryo viability, which already have shown promising results.

5. Conclusion

In conclusion, a novel methodological and automated workflow has been developed to process big data sets of mass spectra generated from MALDI-ToF technology and perform a comparative analysis. The method was demonstrated here to be very fast in comparison with that currently available using freeware software tools for mass spectrometry laboratories [12,13]. Our methodology was also demonstrated to be robust in performing an automatic quality control decision on big data sets up 10000 files.

Author contributions

Ricardo J. Pais drafted the manuscript, performed the analysis and developed python scripts. Ricardo J. Pais and Raminta Zmuidinaite prepared the data sets and generated the R code used for analysis. Ricardo J. Pais, RK Iles, SA Butler and Raminta Zmuidinaite designed the conceptual framework for processing mass spectral data from urines. RK Iles and SA Butler edited the manuscript, managed and funded the project.

Ethics approval

This study was approved by King's College Hospital Ethics Committee (02-03-033). And by the internal Ethics committee of MAP Sciences, where all samples were received anonymised for analysis. Written informed consent was obtained from all women agreeing to participate in the study and held at King's College Hospital.

Conflicts of interest

RK Iles and SA Butler have filed patents on MALDI ToF mass spectral profiling. RK Iles, and SA Butler declare a potential conflict of interest through part ownership of shares in MAP Sciences Ltd. R. Zmuidinaite and, Ricardo J. Pais are employees of MAP Sciences Ltd.

Acknowledgement

Sample collection and patient sample management (anonymisation and meta-data collection and validation) was supported by a grant from the Fetal Medicine Foundation (Charity No: 1037116). This study was part funded by Innovate UK in a grant to MAP IP Holding Ltd, a MAP Sciences Company. Mass spectral Analysis was also funded by MAPSciences Ltd.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2019.100194.

References

[1] Angeletti S. Matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS) in clinical microbiology. J Microbiol Methods 2017;138:20–9. https://doi.org/10.1016/j.mimet.2016.09.003.

[2] Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P, Rolain JM, Raoult D. Ongoing revolution in bacteriology: routine identification of bacteria by matrix--assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 2009;49:543–51. https://doi.org/10.1086/600885.

[3] Nomura F. Proteome-based bacterial identification using matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS): a revolutionary shift in clinical diagnostic microbiology. Biochim Biophys Acta 2015;1854:528–37. https://doi.org/10.1016/j.bbapap.2014.10.022.

[4] Gaillot O, Blondiaux N, Loïez C, Wallet F, Lemaître N, Herwegh S, Courcol RJ. Cost-effectiveness of switch to matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine bacterial identification. J Clin Microbiol 2011;49:4412. https://doi.org/10.1128/JCM.05429-11.

[5] Tan KE, Ellis BC, Lee R, Stamper PD, Zhang SX, Carroll KC. Prospective evaluation of a matrix-assisted laser desorption ionization-time of flight mass spectrometry system in a hospital clinical microbiology laboratory for identification of bacteria and yeasts: a bench-by-bench study for assessing the impact on time to identification and cost-effectiveness. J Clin Microbiol 2012;50:3301–8. https://doi.org/10.1128/JCM.01405-12.

[6] Wang L, Liu HY, Shi HH, Lang JH, Sun W. Urine peptide patterns for non-invasive diagnosis of endometriosis: a preliminary prospective study. Eur J Obstet Gynecol Reprod Biol 2014;177:23–8. https://doi.org/10.1016/j.ejogrb.2014.03.011.

[7] Narasimhan K, Lin SL, Tong T, Baig S, Ho S, Sukumar P, Biswas A, Hahn S, Bajic VB, Choolani M. Maternal serum protein profile and immune response protein subunits as markers for non-invasive prenatal diagnosis of trisomy 21, 18, and 13. Prenat Diagn 2013;33:223–31. https://doi.org/10.1002/pd.4047.

[8] González-Buitrago JM, Ferreira L, Lorenzo I. Urinary proteomics. Clin Chim Acta 2007;375:49–56. https://doi.org/10.1016/j.cca.2006.07.027.

[9] Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P, Dame ZT, Poelzer J, Huynh J, Yallou FS, Psychogios N, Dong E, Bogumil R, Roehring C, Wishart DS. The human urine metabolome. PLoS One 2013;8:e73076https://doi.org/10.1371/journal.pone.0073076.

[10] Duncan MW, Nedelkov D, Walsh R, Hattan SJ. Applications of MALDI mass spectrometry in clinical chemistry. Clin Chem 2015;62:1–10. https://doi.org/10.1373/clinchem.2015.239491.

[11] Sauve AC, T. p.. Speed, Normalization, baseline correction and alignment of high-throughput mass spectrometry data. Proc Gensips 2004:1–4https://www.stat.berkeley.edu/~terry/Group/publications/Final2Gensips2004Sauve.pdf%0Ahttp://stat-www.berkeley.edu/users/terry/Group/publications/Final2Gensips2004Sauve.pdf.

[12] Strohalm M, Kavan D, Novák P, Volný M, Havlícek V. mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data. Anal Chem 2010;82:4648–51. https://doi.org/10.1021/ac100818g.

[13] Strohalm M, Hassman M, Kosata B, Kodícek M. mMass data miner: an open source alternative for mass spectrometric data analysis. Rapid Commun Mass Spectrom 2008;22:905–8. https://doi.org/10.1002/rcm.3444.

[14] Poon LCY, Kametas N, Bonino S, Vercellotti E, Nicolaides KH. Urine albumin concentration and albumin-to-creatinine ratio at 11(+0) to 13(+6) weeks in the prediction of pre-eclampsia. BJOG 2008;115:866–73. https://doi.org/10.1111/j.1471-0528.2007.01650.x.

[15] Iles RK, Cole LA, Butler SA. Direct analysis of hCGβcf glycosylation in normal and aberrant pregnancy by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Int J Mol Sci 2014;15:10067–82. https://doi.org/10.3390/ijms150610067.

[16] Lee SM, Park JS, Norwitz ER, Kim SM, Kim BJ, Park C-W, Jun JK, Syn HC. Characterization of discriminatory urinary proteomic biomarkers for severe pre-eclampsia using SELDI-TOF mass spectrometry. J Perinat Med 2011;39:391–6. https://doi.org/10.1515/jpm.2011.028.

[17] Buhimschi IA, Zhao G, Funai EF, Harris N, Sasson IE, Bernstein IM, Saade GR, Buhimschi CS. Proteomic profiling of urine identifies specific fragments of SERPINA1 and albumin as biomarkers of preeclampsia. Am J Obstet Gynecol 2008;199:551. https://doi.org/10.1016/j.ajog.2008.07.006. e1–16.

[18] Iles RK, Shahpari ME, Cuckle H, Butler SA. Direct and rapid mass spectral fingerprinting of maternal urine for the detection of Down syndrome pregnancy. Clin Proteonomics 2015;12:9. https://doi.org/10.1186/s12014-015-9082-9.

[19] Trivedi DK, Iles RK. Do not just do it, do it right: urinary metabolomics–establishing clinically relevant baselines. Biomed Chromatogr 2014;28:1491–501.

[20] Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics 2012;28:2270–1. https://doi.org/10.1093/bioinformatics/bts447.

[21] Galleani L, Cohen L, Nelson D. Local signal to noise ratio. 2006. p. 6313. https://doi.org/10.1117/12.684026. 63130Q–9.

[22] Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 1964;36:1627–39. https://doi.org/10.1021/ac60214a047.

[23] Gorry PA. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. Anal Chem 1990;62:570–3. https://doi.org/10.1021/ac00205a007.

[24] Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal Chem 1991;63:1193A–203A.

[25] Domon B, Aebersold R. Mass spectrometry and protein analysis. Science 2006;312(80):212–7.

[26] Li F, Yu Z, Chen P, Lin G, Li T, Hou L, Du Y, Tan W. The increased excretion of urinary orosomucoid 1 as a useful biomarker for bladder cancer. Am J Cancer Res 2016;6:331.

[27] Flatley B, Wilmott KG, Malone P, Cramer R. MALDI MS profiling of post-DRE urine samples highlights the potential of β-microseminoprotein as a marker for prostatic diseases. Prostate 2014;74:103–11.

[28] Chen C-J, Liao W-L, Chang C-T, Liao H-Y, Tsai F-J. Urine proteome analysis by C18 plate–matrix-assisted laser desorption/ionization time-of-flight mass spectrometry allows noninvasive differential diagnosis and prediction of diabetic nephropathy. PLoS One 2018;13:e0200945.

[29] Iles RK, Lee CL, Howes I, Davies S, Edwards R, Chard T. Immunoreactive β-core-like material in normal postmenopausal urine: human chorionic gonadotrophin or LH origin? Evidence for the existence of LH core. J Endocrinol 1992;133:459–66.

[30] Lee CL, Iles RK, Shepherd JH, Hudson CN, Chard T. The purification and development of a radioimmunoassay for beta-core fragment of human chorionic gonadotrophin in urine: application as a marker of gynaecological cancer in premenopausal and postmenopausal women. J Endocrinol 1991;130:481–9http://www.ncbi.nlm.nih.gov/pubmed/1719119, Accessed date: 21 May 2019.

[31] Pan S, Zhang H, Rush J, Eng J, Zhang N, Patterson D, Comb MJ, Aebersold R. High throughput proteome screening for biomarker detection*. Mol Cell Proteom 2005;4:182–90. https://doi.org/10.1074/mcp.M400161-MCP200.

[32] Haslam C, Hellicar J, Dunn A, Fuetterer A, Hardy N, Marshall P, Paape R, Pemberton M, Resemannand A, Leveridge M. The evolution of MALDI-TOF mass spectrometry toward ultra-high-throughput screening: 1536-well format and beyond. J Biomol Screen 2016;21:176–86. https://doi.org/10.1177/1087057115608605.

[33] Weisser H, Nahnsen S, Grossmann J, Nilse L, Quandt A, Brauer H, Sturm M, Kenar E, Kohlbacher O, Aebersold R, Malmström L. An automated pipeline for high-throughput label-free quantitative proteomics. J Proteome Res 2013;12:1628–44. https://doi.org/10.1021/pr300992u.

[34] Malm EK, Srivastava V, Sundqvist G, Bulone V. APP: an Automated Proteomics Pipeline for the analysis of mass spectrometry data based on multiple open access tools. BMC Bioinf 2014;15:1. https://doi.org/10.1186/s12859-014-0441-8.

[35] Ao Kong A, Gupta C, Ferrari M, Agostini M, Bedin C, Bouamrani A, Tasciotti E, Azencott R. Biomarker signature discovery from mass spectrometry data. IEEE ACM Trans Comput Biol Bioinform 2014;11:766–72. https://doi.org/10.1109/TCBB.2014.2318718.